Incorporating Conditionally Representative Auxiliary Information in Data Fusion

Maria DeYoreo

Department of Statistical Science
Duke University

Joint work with Bailey Fosdick and Jerry Reiter

Research supported by the National Science Foundation under award SES-11-31897

NCRN Virtual Seminar, October 7, 2015

General Data Fusion Framework

- In many applications, analysts seek to combine two or more databases containing information on disjoint sets of individuals and distinct sets of variables
- Why?
  - Single-source data difficult to obtain due to limited resources (e.g., time, money, or participation)
  - Availability of data varies across sources (e.g., behavior versus opinion)

	A	B	B′
Survey 1	✓	✓	?
Survey 2	✓	?	✓

General Data Fusion Framework

- In many applications, analysts seek to combine two or more databases containing information on disjoint sets of individuals and distinct sets of variables
- Why?
  - Single-source data difficult to obtain due to limited resources (e.g., time, money, or participation)
  - Availability of data varies across sources (e.g., behavior versus opinion)

	A	B	B′
Survey 1	✓	✓	?
Survey 2	✓	?	✓

Applications

- Examples:
  - Marketing: purchasing habits and media viewing habits, e.g., products one purchases and television channels one watches (Gilula et al., 2006)
  - Business: customer satisfaction with bank staff and measures of importance of the customer monetarily to the bank such as funds, number of transactions (Kamakura and Wedel, 1997)
  - Health: cigarette smoking status and opinions about smoking in public (Gilula and McCulloch, 2013)
  - Government and economics: combining microdata from sample surveys (Moriarty and Scheuren, 2003)

File Concatenation

- The problem occurs whenever a researcher needs to consolidate results obtained from two independent samples
- Rubin (1986) emphasizes that data fusion, or file concatenation, can be cast as a missing data problem
- Missing data mechanism is deterministic, i.e., ignorable
- Early work (1980s and 1990s) focused on continuous variables
- Frequently in applications, all variables are categorical

File Concatenation

- The problem occurs whenever a researcher needs to consolidate results obtained from two independent samples
- Rubin (1986) emphasizes that data fusion, or file concatenation, can be cast as a missing data problem
- Missing data mechanism is deterministic, i.e., ignorable
- Early work (1980s and 1990s) focused on continuous variables
- Frequently in applications, all variables are categorical

Common Categorical Data Fusion Methods

- Statistical matching or hot-deck procedures based on A
  - Hot-deck procedures duplicate data on basis of some heuristic
  - Missing values from sample 1 (recipient) are replaced with values from sample 2 (donor)
  - If variables are quantitative, match based on some distance function
  - Perfect matching, perhaps based on subset of "critical variables"
  - Form disjoint clusters (imputation groups) based on A
- Model-based procedures
  - Estimate regression models $P(B \mid A)$ and $P(B' \mid A)$, and use these to predict missing B and B'
  - Estimate models for the joint $P(A, B, B')$
    - Multinomial distribution with log-linear model constraints
    - Latent-class model (Kamakura and Wedel, 1997)

Common Categorical Data Fusion Methods

- Statistical matching or hot-deck procedures based on A
  - Hot-deck procedures duplicate data on basis of some heuristic
  - Missing values from sample 1 (recipient) are replaced with values from sample 2 (donor)
  - If variables are quantitative, match based on some distance function
  - Perfect matching, perhaps based on subset of "critical variables"
  - Form disjoint clusters (imputation groups) based on A
- Model-based procedures
  - Estimate regression models $P(B \mid A)$ and $P(B' \mid A)$, and use these to predict missing B and B'
  - Estimate models for the joint $P(A, B, B')$
    - Multinomial distribution with log-linear model constraints
    - Latent-class model (Kamakura and Wedel, 1997)

Identification Problem

- Goals:
  - ⋆ Fuse databases D_1 on $\{A, B\}$ and D_2 on $\{A, B'\}$ to make inference on distributional quantities, functionals of $P(A, B, B')$
  - ⋆ Generate complete data files that are representative of the population
- $\{A, B, B'\}$ never observed simultaneously $\rightarrow P(A, B, B')$ not identifiable based on D_1 and D_2 alone
- Marginals $P(A, B)$ and $P(A, B')$ constrain $P(A, B, B')$, but many possible specifications of the joint may be consistent with the observed marginal distributions
- The data provide no information on which specifications to favor!

Identification Problem

- Goals:
  - ⋆ Fuse databases D_1 on $\{A, B\}$ and D_2 on $\{A, B'\}$ to make inference on distributional quantities, functionals of $P(A, B, B')$
  - ⋆ Generate complete data files that are representative of the population
- $\{A, B, B'\}$ never observed simultaneously $\rightarrow P(A, B, B')$ not identifiable based on D_1 and D_2 alone
- Marginals $P(A, B)$ and $P(A, B')$ constrain $P(A, B, B')$, but many possible specifications of the joint may be consistent with the observed marginal distributions
- The data provide no information on which specifications to favor!

Typical Assumptions

- Generally proceed by making strong and unverifiable assumptions
- Standard (implicit) assumption: B **and** B' **are independent given** A
- Reasonableness of this assumption depends on richness of A variables and $\{B, B'\}$ dependence
- Ex: every person with the same age, gender, race has the same probability of purchasing an apple computer regardless of media viewing habits
- In some demographics groups, those who do not see ads due to lack of TV/Internet activity may be less likely to purchase product

Typical Assumptions

- Generally proceed by making strong and unverifiable assumptions
- Standard (implicit) assumption: B **and** B' **are independent given** A
- Reasonableness of this assumption depends on richness of A variables and $\{B, B'\}$ dependence
- Ex: every person with the same age, gender, race has the same probability of purchasing an apple computer regardless of media viewing habits
- In some demographics groups, those who do not see ads due to lack of TV/Internet activity may be less likely to purchase product

Relaxing Conditional Independence

Previous work relaxing this assumption:

- Rubin (1986) proposes a sensitivity analysis to values of the partial correlation
- Gilula et al. (2006) propose adding information through a prior on the partial correlation when B and B' are binary
- Gilula and McCulloch (2013) extend this approach to handle variables with more than 2 categories

Our Approach to Relaxing Conditional Independence

Technological advances in recent decades create new exciting opportunities for survey administration.

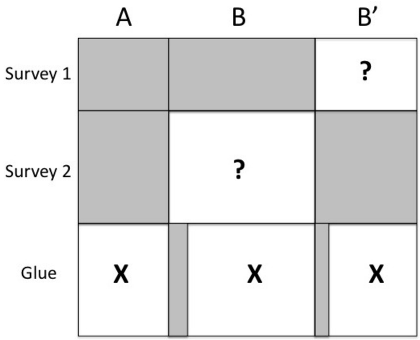
We consider a situation where auxiliary information, i.e. **glue**, is available or obtainable.

	A	B	B’
Survey 1	✓	✓	?
Survey 2	✓	?	✓
Glue			

Glue

We make two assumptions about the glue:

- Represented as additional observations on subsets of $\{A, B, B'\}$
- Each glue observation contains at least one variable in B and one variable in B'



Glue Types

Possible types of glue:

1. Select pairs of variables, e.g. B_i and B'_j
2. All B and B' variables
3. All B and B' variables, some A variables
4. All $\{A, B, B'\}$ variables

How can analysts leverage information in these supplementary surveys for more accurate fusion?

Glue Types

Possible types of glue:

1. Select pairs of variables, e.g. B_i and B'_j
2. All B and B' variables
3. All B and B' variables, some A variables
4. All $\{A, B, B'\}$ variables

How can analysts leverage information in these supplementary surveys for more accurate fusion?

HarperCollins Publishers

- HarperCollins Publishers is one of the world's largest publishing companies
- Contracts research agencies to use stratified sampling procedures to survey people's book buying and reading habits in each country
- Surveys of U.S. population: Pilot (book discovery), Adult (author readership), Product (product utilization), ...
- Each survey consists of
  - basic demographic and reading questions
  - survey specific questions
- All variables are categorical; a common feature in data fusion applications

HarperCollins Publishers

- HarperCollins Publishers is one of the world's largest publishing companies
- Contracts research agencies to use stratified sampling procedures to survey people's book buying and reading habits in each country
- Surveys of U.S. population: Pilot (book discovery), Adult (author readership), Product (product utilization), ...
- Each survey consists of
  - basic demographic and reading questions
  - survey specific questions
- All variables are categorical; a common feature in data fusion applications

HarperCollins Fusion Problem

HarperCollins is interested in understanding the relationship between

1. How an individual becomes aware of an author or book (Pilot survey)
    - on Best Seller List?
    - through Facebook?
    - seeing the book/author's name in a library?
    - ...
2. Which authors an individual prefers (Adult survey)
    - Stephenie Meyer
    - Suzanne Collins
    - Agatha Christie
    - ...

Goal: Combine information in Pilot and Adult surveys to make inference

HarperCollins Fusion Problem

HarperCollins is interested in understanding the relationship between

1. How an individual becomes aware of an author or book (Pilot survey)
   - on Best Seller List?
   - through Facebook?
   - seeing the book/author's name in a library?
   - ...
2. Which authors an individual prefers (Adult survey)
   - Stephenie Meyer
   - Suzanne Collins
   - Agatha Christie
   - ...

Goal: Combine information in Pilot and Adult surveys to make inference

CivicScience

- CivicScience is an Internet polling company that offers real-time insights on public opinion by surveying thousands of people daily
- Surveys are voluntary and available on various internet websites
- Each survey consists of at least the following questions:
  1. Engagement (e.g., Who will win the Superbowl?)
  2. Value (question(s) asked on behalf of paying client)
  3. Profile (demographics)
- Participants may answer additional questions
- Able to connect responses from multiple surveys for some users
- CivicScience was our "glue collector" asking about author readership or discovery (Q2), and gender or age (Q3)

CivicScience

- CivicScience is an Internet polling company that offers real-time insights on public opinion by surveying thousands of people daily
- Surveys are voluntary and available on various internet websites
- Each survey consists of at least the following questions:
  1. Engagement (e.g., Who will win the Superbowl?)
  2. Value (question(s) asked on behalf of paying client)
  3. Profile (demographics)
- Participants may answer additional questions
- Able to connect responses from multiple surveys for some users
- CivicScience was our "glue collector" asking about author readership or discovery (Q2), and gender or age (Q3)

CivicScience

- CivicScience is an Internet polling company that offers real-time insights on public opinion by surveying thousands of people daily
- Surveys are voluntary and available on various internet websites
- Each survey consists of at least the following questions:
  1. Engagement (e.g., Who will win the Superbowl?)
  2. Value (question(s) asked on behalf of paying client)
  3. Profile (demographics)
- Participants may answer additional questions
- Able to connect responses from multiple surveys for some users
- CivicScience was our "glue collector" asking about author readership or discovery (Q2), and gender or age (Q3)

CivicScience

- CivicScience is an Internet polling company that offers real-time insights on public opinion by surveying thousands of people daily
- Surveys are voluntary and available on various internet websites
- Each survey consists of at least the following questions:
  1. Engagement (e.g., Who will win the Superbowl?)
  2. Value (question(s) asked on behalf of paying client)
  3. Profile (demographics)
- Participants may answer additional questions
- Able to connect responses from multiple surveys for some users
- CivicScience was our "glue collector" asking about author readership or discovery (Q2), and gender or age (Q3)

CivicScience

- CivicScience is an Internet polling company that offers real-time insights on public opinion by surveying thousands of people daily
- Surveys are voluntary and available on various internet websites
- Each survey consists of at least the following questions:
  1. Engagement (e.g., Who will win the Superbowl?)
  2. Value (question(s) asked on behalf of paying client)
  3. Profile (demographics)
- Participants may answer additional questions
- Able to connect responses from multiple surveys for some users
- CivicScience was our “glue collector” asking about author readership or discovery (Q2), and gender or age (Q3)

Flexible Bayesian model for multivariate categorical data

- $Y_{ij} \in \{1, \dots, d_j\}$, for $j = 1, \dots, p$, $i = 1, \dots, n$ forms a contingency table with $\prod_{j=1}^{p} d_j$ cells
- Dirichlet process (DP) mixture of product-multinomials (DPM-PM; Dunson and Xing, 2009) for multivariate categorical data

$$Y_{i1}, \dots, Y_{ip} | Z_i, \phi \sim \prod_{j=1}^{p} \text{categorical}(\phi^{(j)}_{z_i,1}, \dots, \phi^{(j)}_{z_i,d_j}),\ i = 1, \dots, n$$

$$\Pr(Z_i = h \mid \pi) = \pi_h,\ i = 1, \dots, n,\ h = 1, \dots, N$$

$$\pi_h = V_h \prod_{g=1}^{h-1} (1 - V_g),\ h = 1, \dots, N$$

$$V_h \sim \text{Beta}(1, \alpha),\ h = 1, \dots, N-1,\ V_N = 1$$

$$\phi^{(j)}_h \sim \text{Dirichlet}(a^{(j)}_1, \dots, a^{(j)}_{d_j}),\ h = 1, \dots, N,\ j = 1, \dots, p$$

$$\alpha \sim \text{gamma}(a_\alpha, b_\alpha)$$

Flexible Bayesian model for multivariate categorical data

- $Y_{ij} \in \{1, \ldots, d_j\}$, for $j = 1, \ldots, p$, $i = 1, \ldots, n$ forms a contingency table with $\prod_{j=1}^{p} d_j$ cells
- Dirichlet process (DP) mixture of product-multinomials (DPM-PM; Dunson and Xing, 2009) for multivariate categorical data

$$Y_{i1}, \ldots, Y_{ip} | Z_i, \phi \sim \prod_{j=1}^{p} \text{categorical}(\phi^{(j)}_{z_i,1}, \ldots, \phi^{(j)}_{z_i,d_j}),\ i = 1, \ldots, n$$

$$\Pr(Z_i = h \mid \pi) = \pi_h,\ i = 1, \ldots, n,\ h = 1, \ldots, N$$

$$\pi_h = V_h \prod_{g=1}^{h-1} (1 - V_g),\ h = 1, \ldots, N$$

$$V_h \sim \text{Beta}(1, \alpha),\ h = 1, \ldots, N-1,\ V_N = 1$$

$$\phi^{(j)}_h \sim \text{Dirichlet}(a^{(j)}_1, \ldots, a^{(j)}_{d_j}),\ h = 1, \ldots, N,\ j = 1, \ldots, p$$

$$\alpha \sim \text{gamma}(a_\alpha, b_\alpha)$$

Properties

- Parsimoniously represents the joint distribution of numerous variables

$$P\left(Y_i = (y_{i1}, \dots, y_{ip}) \mid \pi, \phi\right) = \prod_{k=1}^{N} \pi_k \prod_{j=1}^{p} \phi_{k,y_{ij}}^{(j)}$$

- attractive properties: full support (flexible) and consistent
- computationally tractable
  - No need to determine optimal number of classes, just fix truncation level N large
  - MCMC requires only Gibbs samplers
- Missing Y_{ij} easily imputed during MCMC

Properties

- Parsimoniously represents the joint distribution of numerous variables

$$P\left(Y_i = (y_{i1}, \ldots, y_{ip}) \mid \pi, \phi\right) = \prod_{k=1}^{N} \pi_k \prod_{j=1}^{p} \phi_{k,y_{ij}}^{(j)}$$

- attractive properties: full support (flexible) and consistent
- computationally tractable
  - No need to determine optimal number of classes, just fix truncation level N large
  - MCMC requires only Gibbs samplers
- Missing Y_{ij} easily imputed during MCMC

Properties

- Parsimoniously represents the joint distribution of numerous variables

$$P\left(Y_i = (y_{i1}, \ldots, y_{ip}) \mid \pi, \phi\right) = \prod_{k=1}^{N} \pi_k \prod_{j=1}^{p} \phi_{k, y_{ij}}^{(j)}$$

- attractive properties: full support (flexible) and consistent
- computationally tractable
  - No need to determine optimal number of classes, just fix truncation level N large
  - MCMC requires only Gibbs samplers
- Missing Y_{ij} easily imputed during MCMC

Incorporating Glue

- Databases D_1 of size n_1 on $\{A, B\}$ and D_2 of size n_2 on $\{A, B'\}$
  - Y_{ij} for $j \in A$ is observed for all $n_1 + n_2$ individuals
  - Y_{ij} for $j \in B$ is observed for n_1 individuals in D_1
  - Y_{ij} for $j \in B'$ is observed for n_2 individuals in D_2
  - Item nonresponse $\rightarrow$ missing values within D_1 and D_2 also
- Assume glue D_s of size n_s containing subset of $\{A, B, B'\}$
- Concatenate (D_1, D_2, D_s) in one file and estimate DPM-PM model, in process imputing missing B in D_1 and missing B' in D_2, but not missing values in D_s
- Information on $\{A, B, B'\}$ in D_s influences parameter estimates resulting in imputations for B and B' that reflect dependence in glue

Incorporating Glue

- Databases D_1 of size n_1 on $\{A, B\}$ and D_2 of size n_2 on $\{A, B'\}$
  - Y_{ij} for $j \in A$ is observed for all $n_1 + n_2$ individuals
  - Y_{ij} for $j \in B$ is observed for n_1 individuals in D_1
  - Y_{ij} for $j \in B'$ is observed for n_2 individuals in D_2
  - Item nonresponse $\rightarrow$ missing values within D_1 and D_2 also
- Assume glue D_s of size n_s containing subset of $\{A, B, B'\}$
- Concatenate (D_1, D_2, D_s) in one file and estimate DPM-PM model, in process imputing missing B in D_1 and missing B' in D_2, but not missing values in D_s
- Information on $\{A, B, B'\}$ in D_s influences parameter estimates resulting in imputations for B and B' that reflect dependence in glue

Simulation study: HarperCollins Publishers

Product survey - 3567 respondents

A variables:

- Gender
- Age - { 18-24, 25-34, 35-44, 45-54, 55-64, 65+ }
- Income - { $<$25K, 25-45K, 45-75K, 75-99K, 100+ K, Prefer not say }
- Work status - { emp FT, emp PT, homemaker, retired, self-emp, other }

B variable:

- eBook reader ownership - { yes, no }

B' variable:

- Reading hours per week - {$<$1 hour, 1-4 hours, 5$+$ hours}

Simulation procedure

1. Randomly split data set to create fusion situation with missing B and B'
2. Consider the following glue scenarios:
   - No glue
   - {eBook (B), hours (B')}
   - {eBook (B), hours (B'), gender (A_g)}
   - {eBook (B), hours (B'), age (A_a)}
   - {eBook (B), hours (B'), gender (A_g), age (A_a)}

   Glue contains the observed variables for all survey respondents
3. Estimate the DPM-PM model using MCMC
4. Obtain 120,000 samples of parameters saving 50 complete (imputed) data files

Simulation procedure

1. Randomly split data set to create fusion situation with missing B and B'
2. Consider the following glue scenarios:
   - No glue
   - {eBook (B), hours (B')}
   - {eBook (B), hours (B'), gender (A_g)}
   - {eBook (B), hours (B'), age (A_a)}
   - {eBook (B), hours (B'), gender (A_g), age (A_a)}

   Glue contains the observed variables for all survey respondents
3. Estimate the DPM-PM model using MCMC
4. Obtain 120,000 samples of parameters saving 50 complete (imputed) data files

Quantifying the impact of glue

1. Hellinger distance $2^{-1/2}\sqrt{\sum_{i=1}^{K}(\sqrt{p_i}-\sqrt{q_i})^2}$ between empirical distribution of (A_g, A_a, B, B') based on original complete survey and posterior inferences

Table: Posterior distributions of the Hellinger distances for various glue types. 10 perfect matching data sets considered.

	mean	95% CI or range*
no glue	.104	(.094, .113)
$\{B, B'\}$	.083	(.075, .091)
$\{B, B', A_g\}$	.077	(.071, .084)
$\{B, B', A_a\}$	.060	(.053, .068)
$\{B, B', A_g, A_a\}$	.052	(.047, .059)
Exact matching	.100	.090 - .107

Quantifying the impact of glue

1. Hellinger distance $2^{-1/2}\sqrt{\sum_{i=1}^{K}(\sqrt{p_i}-\sqrt{q_i})^2}$ between empirical distribution of (A_g, A_a, B, B') based on original complete survey and posterior inferences

Table: Posterior distributions of the Hellinger distances for various glue types. 10 perfect matching data sets considered.

	mean	95% CI or range*
no glue	.104	(.094, .113)
$\{B, B'\}$	.083	(.075, .091)
$\{B, B', A_g\}$	.077	(.071, .084)
$\{B, B', A_a\}$	.060	(.053, .068)
$\{B, B', A_g, A_a\}$	.052	(.047, .059)
Exact matching	.100	.090 - .107

Quantifying the impact of glue

2. Discrepancy between empirical imputed contingency table and true contingency table yields expected number of misclassified individuals in imputed data set:

$$\frac{1}{50}\sum_{m=1}^{50}\left(0.5\sum_{j=1}^{\prod_{k=1}^{p} d_k}|n_j - \hat{n}_j^{(m)}|\right)$$

Table: Average number of individuals in incorrect cells of the contingency table over the 50 imputed data files. 10 complete data sets considered for the statistical matching procedure.

	$\frac{1}{2}\mathrm{E}\left(\sum_{j=1}^{k}\|n_j - \hat{n}_j\|\right)$
no glue	318
$\{B, B'\}$	250
$\{B, B', A_g\}$	247
$\{B, B', A_a\}$	199
$\{B, B', A_g, A_a\}$	196
Exact matching	315

Quantifying the impact of glue

2. Discrepancy between empirical imputed contingency table and true contingency table yields expected number of misclassified individuals in imputed data set:

$$\frac{1}{50}\sum_{m=1}^{50}\left(0.5\sum_{j=1}^{\prod_{k=1}^{p} d_k}|n_j - \hat{n}_j^{(m)}|\right)$$

Table: Average number of individuals in incorrect cells of the contingency table over the 50 imputed data files. 10 complete data sets considered for the statistical matching procedure.

	$\frac{1}{2}\mathrm{E}\left(\sum_{j=1}^{k}\lvert n_j - \hat{n}_j\rvert\right)$
no glue	318
$\{B, B'\}$	250
$\{B, B', A_g\}$	247
$\{B, B', A_a\}$	199
$\{B, B', A_g, A_a\}$	196
Exact matching	315

Quantifying the impact of glue

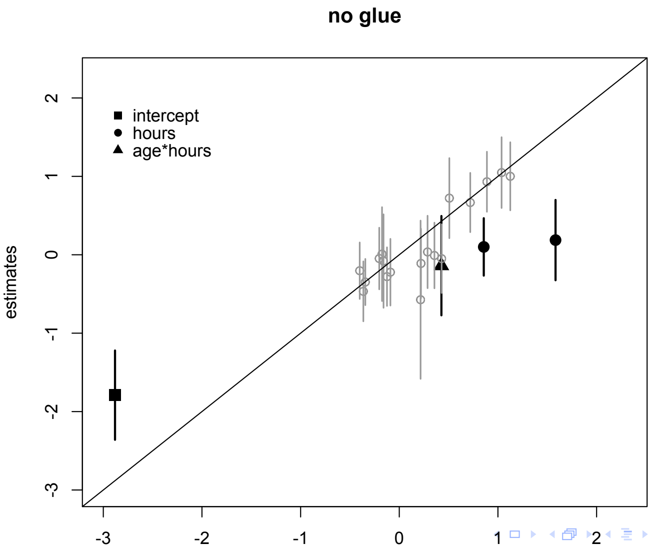
3. Logistic regression model coefficients

$$
\begin{aligned}
\text{logit}(p(eBook = 1)) = \beta_0 + \beta^g 1(gender = female) + \sum_{k=2}^{6} \beta_k^a 1(age = k) \\
+ \sum_{k=2}^{6} \beta_k^w 1(work = k) + \sum_{k=2}^{6} \beta_k^i 1(income = k) + \sum_{k=2}^{3} \beta_k^h 1(hours = k) \\
+ \beta^{hg} 1(hours = 5+, gender = female) + \beta^{ha} 1(age = 65+, hours = 5+) \\
+ \beta^{hga} 1(age = 65+, hours = 5+, gender = female)
\end{aligned}
$$

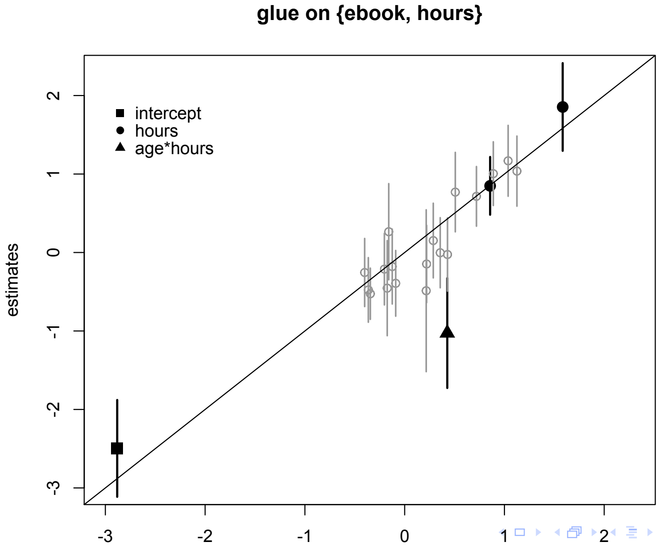
- Compare coefficients estimated by the model with those estimated on the true complete data

- A more focused evaluation

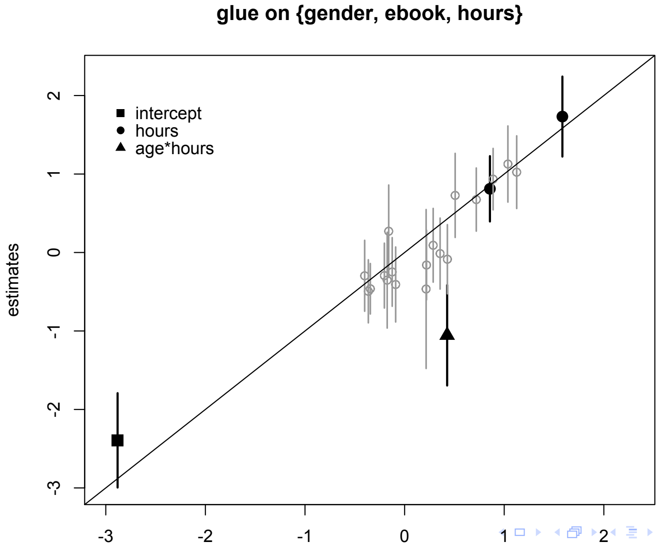
Logistic regression coefficient estimates



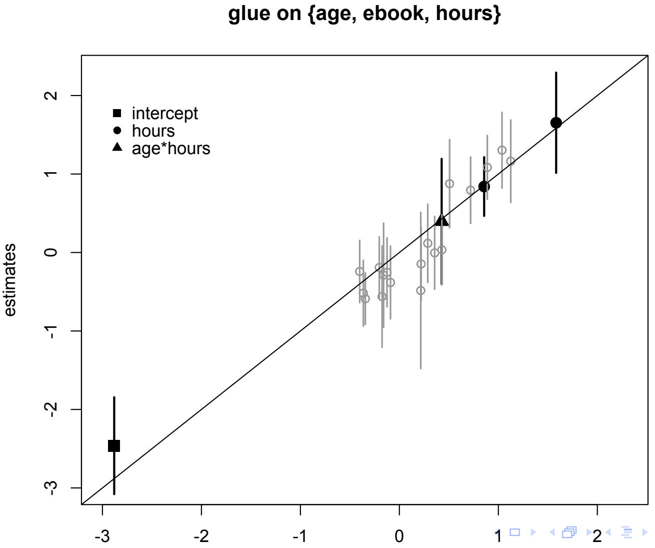
Logistic regression coefficient estimates



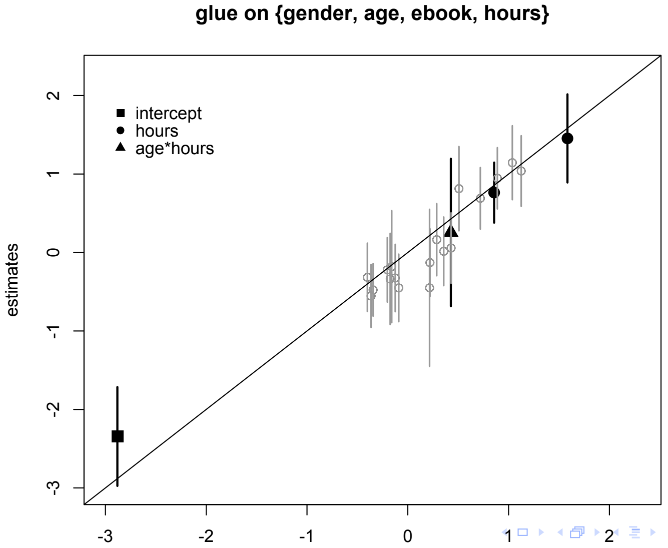
Logistic regression coefficient estimates



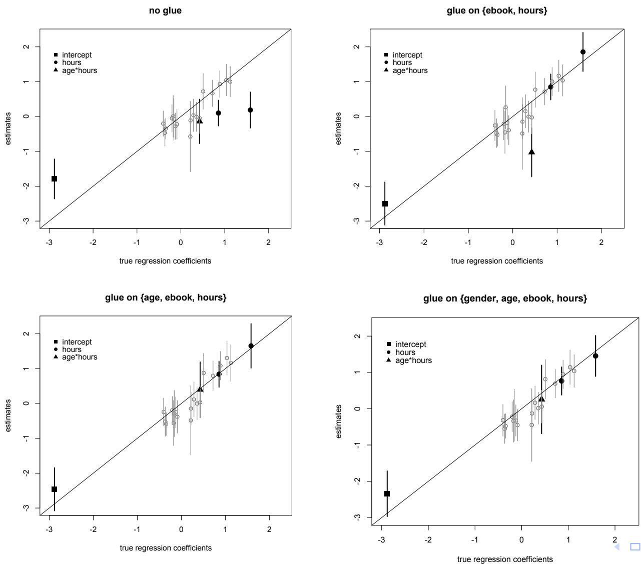
Logistic regression coefficient estimates



Logistic regression coefficient estimates



Logistic regression coefficient estimates



Nonrepresentative Glue

- Voluntary Internet survey
- Over 60% of CivicScience respondents are 55+ compared to only 30% of HarperCollins respondents
- $\{A, B, B'\}$ from supplemental survey data is not representative of the joint from (D_1, D_2)

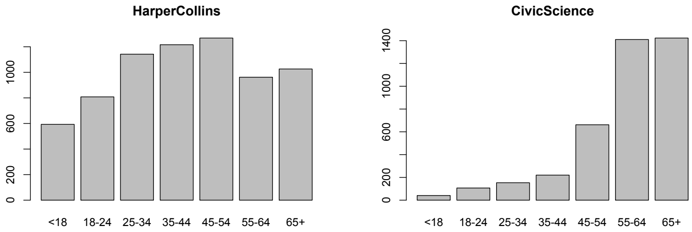


Figure: Age distributions of respondents.

Nonrepresentative Glue

- Voluntary Internet survey
- Over 60% of CivicScience respondents are 55+ compared to only 30% of HarperCollins respondents
- $\{A, B, B'\}$ from supplemental survey data is not representative of the joint from (D_1, D_2)

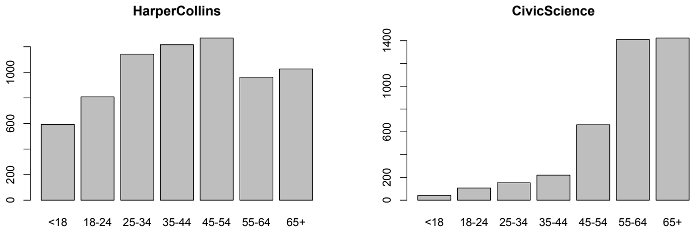


Figure: Age distributions of respondents.

Nonrepresentative Glue

- A common problem to be encountered in practice from convenience or non-probability samples

- Problems can arise even when appending glue that is representative of the population in terms of $P(B, B' \mid A)$ but not on A

Suppose $n_A = n_B = n_{B'} = 1$ and all variables are binary

- Glue collection procedure extremely oversamples subpopulation with $A = 1$, but distribution of $(B, B'|A)$ is representative

- Inference on (B, B') distribution will heavily resemble $P(B, B'|A = 1)$

- If $P(B, B'|A = 1)$ and $P(B, B'|A = 2)$ differ greatly, inference for $P(B, B'|A = 2)$ and $P(B, B')$ will be of poor quality

Nonrepresentative Glue

- A common problem to be encountered in practice from convenience or non-probability samples

- Problems can arise even when appending glue that is representative of the population in terms of $P(B, B' \mid A)$ but not on A

Suppose $n_A = n_B = n_{B'} = 1$ and all variables are binary

- Glue collection procedure extremely oversamples subpopulation with $A = 1$, but distribution of $(B, B'|A)$ is representative

- Inference on (B, B') distribution will heavily resemble $P(B, B'|A = 1)$

- If $P(B, B'|A = 1)$ and $P(B, B'|A = 2)$ differ greatly, inference for $P(B, B'|A = 2)$ and $P(B, B')$ will be of poor quality

Nonrepresentative Glue

- A common problem to be encountered in practice from convenience or non-probability samples

- Problems can arise even when appending glue that is representative of the population in terms of $P(B, B' \mid A)$ but not on A

Suppose $n_A = n_B = n_{B'} = 1$ and all variables are binary

- Glue collection procedure extremely oversamples subpopulation with $A = 1$, but distribution of $(B, B'|A)$ is representative

- Inference on (B, B') distribution will heavily resemble $P(B, B'|A = 1)$

- If $P(B, B'|A = 1)$ and $P(B, B'|A = 2)$ differ greatly, inference for $P(B, B'|A = 2)$ and $P(B, B')$ will be of poor quality

Nonrepresentative Glue

- A common problem to be encountered in practice from convenience or non-probability samples

- Problems can arise even when appending glue that is representative of the population in terms of $P(B, B' \mid A)$ but not on A

Suppose $n_A = n_B = n_{B'} = 1$ and all variables are binary

- Glue collection procedure extremely oversamples subpopulation with $A = 1$, but distribution of $(B, B'|A)$ is representative

- Inference on (B, B') distribution will heavily resemble $P(B, B'|A = 1)$

- If $P(B, B'|A = 1)$ and $P(B, B'|A = 2)$ differ greatly, inference for $P(B, B'|A = 2)$ and $P(B, B')$ will be of poor quality

Incorporating nonrepresentative glue

We propose generating representative glue and then fitting the DPM-PM model with the generated glue to obtain imputations and parameter estimates.

Procedure for generating representative glue:

1. Fit the DPMPM model to the supplementary data alone to estimate $P(A, B, B')$, from which one can obtain $P(B|A, B')$ and $P(B'|A, B)$.
2. Sample records with replacement from databases (D_1 and D_2). Impute missing B' by sampling from $P(B'|A, B)$ and impute missing B by sampling from $P(B|A, B')$ estimated in (1).

Assessing validity of this procedure:

- This assumes the glue is representative of $P(B|A, B')$ and $P(B'|A, B)$
- Evaluate these assumptions by comparing the empirical $P(B)$ and $P(B')$ sampled in step (2) to those in the surveys

Incorporating nonrepresentative glue

We propose generating representative glue and then fitting the DPM-PM model with the generated glue to obtain imputations and parameter estimates.

Procedure for generating representative glue:

1. Fit the DPMPM model to the supplementary data alone to estimate $P(A, B, B')$, from which one can obtain $P(B|A, B')$ and $P(B'|A, B)$.
2. Sample records with replacement from databases (D_1 and D_2). Impute missing B' by sampling from $P(B'|A, B)$ and impute missing B by sampling from $P(B|A, B')$ estimated in (1).

Assessing validity of this procedure:

- This assumes the glue is representative of $P(B|A, B')$ and $P(B'|A, B)$
- Evaluate these assumptions by comparing the empirical $P(B)$ and $P(B')$ sampled in step (2) to those in the surveys

Incorporating nonrepresentative glue

We propose generating representative glue and then fitting the DPM-PM model with the generated glue to obtain imputations and parameter estimates.

Procedure for generating representative glue:

1. Fit the DPMPM model to the supplementary data alone to estimate $P(A, B, B')$, from which one can obtain $P(B|A, B')$ and $P(B'|A, B)$.
2. Sample records with replacement from databases (D_1 and D_2). Impute missing B' by sampling from $P(B'|A, B)$ and impute missing B by sampling from $P(B|A, B')$ estimated in (1).

Assessing validity of this procedure:

- This assumes the glue is representative of $P(B|A, B')$ and $P(B'|A, B)$
- Evaluate these assumptions by comparing the empirical $P(B)$ and $P(B')$ sampled in step (2) to those in the surveys

Motivating question

HarperCollins is interested in understanding the relationship between

$\boldsymbol{B}$: how an individual becomes aware of an author or book (Pilot survey)

- 6 discovery mediums (e.g Facebook, Best Seller List)

$\boldsymbol{B'}$: which authors an individual prefers (Adult survey)

- 5 authors (e.g. Agatha Christie, Stephenie Meyer)

We aim to combine information from the Pilot ($n_1 = 2,000$) and Adult ($n = 5,015$) surveys to address this question.

$\boldsymbol{A}$ variables include gender, age, and income, of interest for market segmentation.

CivicScience glue contains $n_s = 2,730$ observations containing at least one $\{B_j, B'_k\}$ pair.

Motivating question

HarperCollins is interested in understanding the relationship between

***B*:** how an individual becomes aware of an author or book (Pilot survey)

- 6 discovery mediums (e.g Facebook, Best Seller List)

***B′*:** which authors an individual prefers (Adult survey)

- 5 authors (e.g. Agatha Christie, Stephenie Meyer)

We aim to combine information from the Pilot ($n_1 = 2,000$) and Adult ($n = 5,015$) surveys to address this question.

A variables include gender, age, and income, of interest for market segmentation.

CivicScience glue contains $n_s = 2,730$ observations containing at least one $\{B_j, B'_k\}$ pair.

Glue questions - Discovery & Author preferences

B: Do you become aware of new authors through _________?

Answers: Yes, No

1. The Best Seller List
2. Facebook
3. Library
4. Online site
5. Recommendations from friends and family
6. Bookstore

B′: What is your experience with author _________?

Answers: Read, Not read but interested, Not read and not interested

1. Lisa Kleypas (historical and contemporary romance novels)
2. Stephenie Meyer (e.g. Twilight)
3. Suzanne Collins (e.g. The Hunger Games trilogy)
4. Agatha Christie (detective novels and shorty stories)
5. Shel Silverstein (e.g. The Giving Tree)

Generating representative CivicScience glue

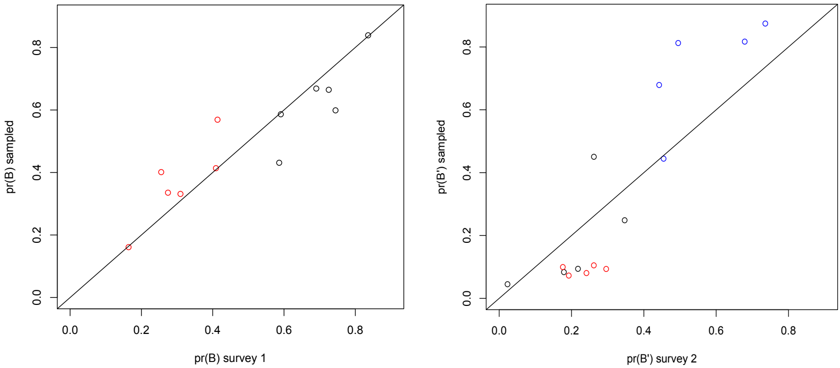


Figure: Left plot: empirical $P(B)$ Right plot: empirical $P(B')$

- Discrepancies evident between sampled $P(B')$ distribution and survey $P(B')$
- Choose to generate glue D_s^* assuming only $P(B|A, B')$ representative

Inference for HarperCollins

- Append the constructed D_s^* to (D_1, D_2) and estimate the DPM-PM model on the concatenated data
- Impute all missing values in D_1 and D_2 in the process
- Completed versions may be used for multiple imputation inference on any functional of $P(A, B, B')$ HarperCollins desires
- Example: probability of discovery via a given medium for those who have read a particular author

Inference for HarperCollins

- Append the constructed D_s^* to (D_1, D_2) and estimate the DPM-PM model on the concatenated data

- Impute all missing values in D_1 and D_2 in the process

- Completed versions may be used for multiple imputation inference on any functional of $P(A, B, B')$ HarperCollins desires

- Example: probability of discovery via a given medium for those who have read a particular author

Discovery Given Readership by Income

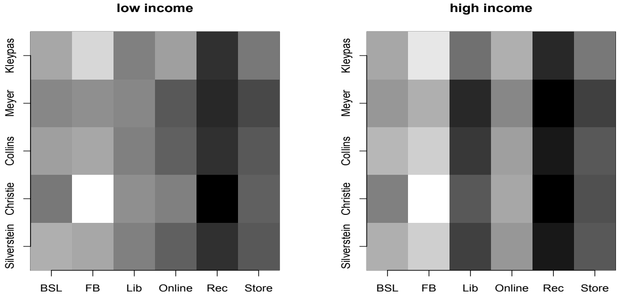


Figure: Point estimates for $\Pr(B = \text{yes} \mid B' = \text{read}, \text{income})$ for low (left) and high (right) income groups for all mediums and authors.

- Among individuals who have read Meyer, those with high incomes are very likely to discover books at library, whereas those with low income are not.
- Low income individuals more likely to discover via Internet for all authors except Kleypas.

Discovery Given Readership by Age

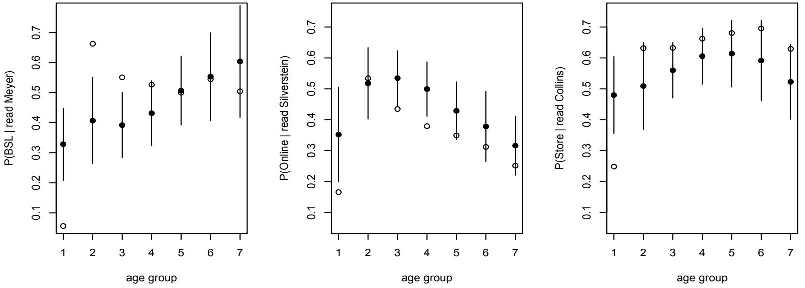


Figure: Estimates for $\Pr(B = \text{yes} \mid B' = \text{read}, \text{age})$ across age for 3 medium/author combinations. Open circles indicate "no glue" estimates.

- Of individuals who have read Meyer, older individuals more likely to discover through BSL
- Estimates without glue agree on trends sometimes (e.g., middle figure), but often very different (left figure)

Remaining Questions

- Simulations point to need for cost-benefit analysis to guide glue collection
- Cost of collecting glue increases with number of variables and observations
- Research on methods for selecting variables that improve the accuracy of data fusion taking into account cost of variables
- Computational improvements – HarperCollins (and other companies) would love if we could fuse **all** of their surveys on hundreds or thousands of variables
- Come up with a better way to use information provided by glue that does not involve fitting MCMC twice and avoids copying observations from the surveys to form the representative glue (⋆Current work⋆)

Remaining Questions

- Simulations point to need for cost-benefit analysis to guide glue collection
- Cost of collecting glue increases with number of variables and observations
- Research on methods for selecting variables that improve the accuracy of data fusion taking into account cost of variables
- Computational improvements – HarperCollins (and other companies) would love if we could fuse **all** of their surveys on hundreds or thousands of variables
- Come up with a better way to use information provided by glue that does not involve fitting MCMC twice and avoids copying observations from the surveys to form the representative glue (⋆Current work⋆)

Thank you!

- Coauthors: Bailey Fosdick (CSU) and Jerry Reiter (Duke)
- Working group members from SAMSI program on Computational Methods in Social Sciences, 2013-2014
- HarperCollins Publishers
- CivicScience
- Thanks to the NCRN and the audience for your attention!

Interested in This Topic?

- Email me: `maria.deyoreo@stat.duke.edu`

- Manuscript available on `arXiv` or my webpage `https://stat.duke.edu/~mnd13/`